

Global AI governance: barriers and pathways forward

Huw Roberts^{1*}, Emmie Hine², Mariarosaria Taddeo,^{1,3} Luciano Floridi^{2,4}

¹ Oxford Internet Institute, University of Oxford, 1 St Giles', Oxford, OX1 3JS, UK

² Department of Legal Studies, University of Bologna, Via Zamboni 22, 40126, Bologna, Italy

³ The Alan Turing Institute, 96 Euston Rd. London, NW1 2DB, UK

⁴ Digital Ethics Center, Yale University, 85 Trumbull Street, New Haven, CT 06511, US

*Email for correspondence: huw.roberts@oii.ox.ac.uk

Abstract

Recent advances in artificial intelligence (AI), combined with a proliferation in use, have led to a newfound emphasis on strengthening the global governance of AI. In this article, we assess the prospects for this aspiration and potential pathways forward. We outline the key global AI governance processes taking place, analyse how first- and second-order cooperation problems in international relations apply to the global governance of AI, and consider how to enhance international cooperation. We argue that a promising foundation of international regimes focused on AI governance is emerging but that the centrality of AI to interstate competition, dysfunctional international institutions, and a lack of agreement over international policy priorities for AI threaten prospects for substantive cooperation. Drawing on lessons from adjacent policy areas, particularly climate governance, we propose strengthening the existing weak “regime complex” of international institutions as the most desirable and realistic path forward for global AI governance. Improving the coordination, capacities, and legitimacy of existing international institutions governing AI would support mutually reinforcing policy change that can address a range of risks associated with these technologies in the most politically legitimate way. We conclude by offering recommendations on how to strengthen the AI regime complex.

Keywords: AI governance, ethics, global governance, international cooperation, regime complex.

1. Introduction

Late 2022 and early 2023 saw significant breakthroughs in the commercialisation of foundation models: artificial intelligence (AI) systems designed to have a wide range of possible applications with minimal modification or fine-tuning (Bommasani et al., 2022). Because these systems can be applied to various tasks, there has been rapid adoption by both individual consumers and companies. OpenAI's ChatGPT, a chatbot built on the company's GPT-3.5 and GPT-4 foundation models, has become the fastest-growing consumer internet application ever (Marcus & Reuel, 2023). These models have also been integrated into products across sectors, from Microsoft's search engine, Bing, to the language learning app Duolingo (Malik, 2023).

Foundation models are already increasing economic efficiency and enhancing consumer experiences, yet they also exacerbate many risks associated with task-specific AI (Küspert et al., 2023; Roberts, Babuta, et al., 2022). They heighten the risk of market concentration, excessive data extraction, and environmentally damaging compute requirements (Bender et al., 2021; Weidinger et al., 2022), while facilitating the creation of fake or malicious information (Helberger & Diakopoulos, 2023; Verma & Oremus, 2023). Many of these risks transcend national borders, including national security threats from democratising capabilities that can be used by malicious actors (Hazell, 2023) and heightened international inequalities from the concentration of high value-adding tasks in specific geographies (Weber, 2017).

The commercialisation of foundation models has reinvigorated discussions about the global governance of AI,¹ understood as the process through which diverse interests that transcend borders are accommodated, without a single sovereign authority, so that cooperative action may be taken in maximising the benefits and mitigating the risks of AI (Finkelstein, 1995; Weiss, 2000). The United Nations (UN) Secretary General Antonio Guterres, British Prime Minister Rishi Sunak, and OpenAI CEO Sam Altman have all argued for the creation of a new international AI body (10 Downing Street, 2023; Altman et al., 2023; UN Press, 2023). Proposals for what a new international AI body could look like typically draw on existing bodies like the International Panel on Climate Change (IPCC) and the International Atomic Energy Agencies (IAEA) for inspiration (Bak-Coleman et al., 2023; Chowdhury, 2023; Ho et al., 2023; Marcus & Reuel, 2023; Samson, 2023). This newfound emphasis on global AI governance is a promising development. However, current proposals lack a strong contextualisation in terms of the current international AI governance landscape and the potential geopolitical and institutional barriers that new initiatives may face. This raises questions about the desirability and feasibility of the proposals (Drexel & Depp, 2023).

In this article, we address this literature gap by assessing the prospects for strengthening global AI governance in light of current geopolitical and institutional realities and considering potential pathways forward. We argue that a promising foundation of international regimes focused on AI governance is emerging but that the centrality of AI to interstate competition, dysfunctional international institutions, and a lack of agreement over international policy priorities for AI threaten prospects for substantive cooperation. Given these conditions, we propose

¹ Prior to 2023, there was a small literature considering global governance mechanisms for AI, for instance through arms control initiatives (Maas, 2019; Scharre & Lamberth, 2022), an international coordinating committee (Jelinek et al., 2021), and private governance mechanisms (Cihon, 2019).

strengthening the existing weak “regime complex” of international institutions as the most desirable and realistic path forward for global AI governance. Strengthening the capacities, coordination, and legitimacy of existing institutions allows for mutually reinforcing policy change, which, if enacted properly, can lead to catalytic change across the various international policy areas which AI impacts and facilitate flexible governance necessary for rapidly evolving technologies.

To make this argument, we outline key global AI governance processes in section two. In section three, we analyse how global governance “gridlock” from first- and second-order cooperation problems in international relations applies to AI (Hale et al., 2013). In section four, we assess possible pathways forward for global AI governance based on the degree to which they are politically legitimate. We draw on lessons from other policy areas, particularly international climate governance, to argue that strengthening the current weak regime complex is the most desirable and realistic option for politically legitimate global governance in the near term. We conclude the article by offering recommendations for strengthening the political legitimacy of the existing weak regime complex.

2. The landscape of global governance initiatives

At the time of writing, at least 44 countries have released national AI strategies (Fatima et al., 2021), with China, the EU, and the US, amongst others, outlining an ambition to strengthen global AI governance (Roberts, Cowls, Hine, et al., 2021; Webster et al., 2017). While it is reasonable to be sceptical about these claims – particularly given concurrent narratives surrounding AI competition (Scharre, 2021) – early efforts have been made to turn them into practice, which we map below.²

2.1. Intergovernmental initiatives

Since 2016, the Organisation for Economic Cooperation and Development (OECD) has been undertaking technocratic work through networks of experts to produce analysis and best practices for AI.³ In May 2019, OECD member countries adopted a set of AI Principles “designed to promote use of AI that is innovative and trustworthy and respects human rights and democratic values” (*The OECD AI Principles*, 2019). G20 leaders subsequently committed to a set of non-binding AI principles that were drawn from the OECD’s (*G20 Osaka Leaders’ Declaration*, 2019).

In November 2021, 193 Member States adopted a UNESCO Recommendation on the Ethics of Artificial Intelligence designed to guide signatories in developing appropriate legal frameworks (*UNESCO AI Ethics Recommendation*, 2022). The UNESCO AI Recommendation outlines principles substantively like the OECD’s, yet the OECD’s principles were not mentioned in this document. Given that the OECD and G20 countries, which account for the vast majority of the world’s AI development, already agreed to the OECD AI ethics principles, the failure to acknowledge existing work is unconstructive and a waste of resources.

In 2023, the G7 jointly called for a “Hiroshima AI Process” to enhance cooperation in AI governance (*Hiroshima Communiqué*, 2023), while the UK announced that it will host a global AI summit in late 2023 focused on AI safety (10 Downing Street, 2023).

² This mapping provides an overview of key initiatives but is not meant to be exhaustive. For a more complete mapping, see (Schmitt, 2022; Veale et al., 2023).

³ <https://oecd.ai/en/about>

Efforts have also been made to establish new AI governance bodies. The Global Partnership on AI (GPAI), conceived within the architecture of the G7 before being launched in 2020 by 15 founding members,⁴ is one example (Schmitt, 2022). The GPAI is founded on the OECD's AI Principles and aims to "support and guide the ethical adoption of AI" through facilitating international cooperation and undertaking practical projects, like developing metrics (*GPAI Terms of Reference*, 2020). However, the GPAI's impact has thus far been limited. Despite most cutting-edge breakthroughs taking place in private sector companies, GPAI expert groups are almost entirely made up of academics (*Multistakeholder Experts Group Report*, 2022), with it questionable whether they will have sufficient access and information to provide appropriate guidance. Moreover, GPAI has been politicised since its outset, with the US only agreeing to join this organisation to counter the influence of China (Schmitt, 2022), suggesting its work may have a limited global influence.

The EU-US Trade and Technology Council (TTC), established in 2021, is a second new intergovernmental body. It serves as a forum for periodic, high-level discussion and coordination between the EU and US governments in areas related to trade and technology. This includes AI, with the TTC publishing a joint roadmap for developing common tools and standards for trustworthy AI in December 2022 (TTC, n.d.). Finally, the UN Tech Envoy also announced that they will establish a High-Level Expert Body on AI tasked with advancing recommendations for the international governance of AI.⁵

There has also been defence-specific cooperation, including the UN-led process of establishing norms for lethal autonomous weapons (LAWs), a new Partnership for Defence, constituted of the US and several of its allies (Stanley-Lockman & Christie, 2021), and ongoing work in NATO, which published an AI strategy (NATO, 2021) and in 2023 established a Data and Artificial Intelligence Review Board which works to operationalise the Organisation's principles for responsible AI via a set of standards to inform the entire AI lifecycle in defence.

An exception to non-binding governance initiatives is work by the Council of Europe (COE).⁶ In September 2019, the COE established an Ad-Hoc Committee on AI (CAHAI) tasked with examining the feasibility of developing an international legal framework for AI, with current discussions focused on creating a legally binding international Convention for AI (Schmitt, 2022). Although the COE comprises 46 mainly European states, non-members, such as the US and Israel, which are involved in negotiations, would be able to sign and ratify the treaty. The current draft contains several obligations focused on ensuring that human rights are respected in the development and deployment of AI systems. Importantly, signatories who ratify this treaty would implement the Convention through domestic laws, meaning individuals would need to seek remedies for breaches of treaty protections at a national level (*COE Rules*, 2023).

2.2. *Private governance initiatives*

In parallel with intergovernmental initiatives, there has been a concerted effort by private stakeholders to deepen cooperation. International technical standards bodies, including but not limited to the International Organization for Standardization (ISO) and International

⁴ There are now 29 GPAI members.

⁵ <https://www.un.org/techenvoy/content/artificial-intelligence>

⁶ Note: the COE is not an EU institution and has a separate membership.

Electrotechnical Commission (IEC) have been working on design and process consistency for AI among countries (Cihon, 2019), some of which have already been published (McGarr, 2023). These international standards are voluntary, yet some may be mandated as *de jure* legal requirements in relevant legislation or become an industry best practice standard and require *de facto* compliance.

Industry has also created new governance institutions. The Partnership on AI (PAI) was established in 2016 as a non-profit organisation comprising “Big Tech” companies, which predominantly fund the organisation, and civil society and academic stakeholders. PAI convenes stakeholders to develop guidance, inform public policy, and enhance public understanding.⁷ However, out of 105 partner organisations, there are eight in Europe, six in Asia, and just two each in Africa and Oceania. By comparison, there are seventeen in Silicon Valley, raising doubts about the representativeness of the organisation.⁸ Although there is evidence of some Big Tech companies following PAI guidance (PAI Staff, 2022), industry self-regulation raises a “principal moral hazard” due to misaligned incentives for this body to produce stringent rules or guidance (Miller & Whitford, 2007). Indeed, the digital rights civil society organisation Access Now resigned from the PAI after concluding the organisation failed to influence the attitude of member companies (Veale et al., 2023). The 2023 Frontier Model Forum, founded by four Big Tech companies to promote good governance of cutting-edge systems, faces the same ethical risks.⁹

There have also been *ad hoc* transnational initiatives from non-profit organisations designed to influence global rules for AI. The most well-known is the Future of Life Institute-led open letter to implement a six-month pause on the development of AI systems more advanced than OpenAI’s GPT-4 over fears of an unsafe race dynamic (“Pause Giant AI Experiments” 2023). Despite receiving significant criticism for drawing attention away from harms already materialising, it has influenced international policy initiatives, with the announcement of the UK’s AI Summit directly citing the letter (“AI Poses Risks Today,” 2023). Similar letters have been produced by coalitions of civil society organisations focused on lobbying the EU to strengthen its rules, most recently in relation to foundation models (Kak et al., 2023). Notably, many of these calls targeting EU policy come from organisations and individuals outside the bloc, highlighting the extraterritorial influence EU unilateral rulemaking could have (Siegmann & Anderljung, 2022).

3. Barriers to strong global AI governance

States build international institutions – “regimes” – to facilitate cooperation. However, there is not a single model for regime structure. Some policy areas are integrated and hierarchical, while others are fragmented with no identifiable core and weak linkages between nodes. Different institutional arrangements sit between these ends, including semi-hierarchical regimes and non-hierarchical, loosely tied “regime complexes” (Keohane & Victor, 2011).

The global AI governance landscape is best understood as a weak regime complex, as it has a “polycentric” structure with some linkages between institutions – such as the G7’s Hiroshima Process drawing on work from the OECD and GPAI – but work is generally siloed (Cihon et al., 2021; Tallberg et al., 2023). For some aspects of AI policy, this is relatively unproblematic. AI is not a single policy problem, but rather a set of loosely connected problems that arise from

⁷ <https://partnershiponai.org/about/>

⁸ <https://partnershiponai.org/partners/>

⁹ <https://blog.google/outreach-initiatives/public-policy/google-microsoft-openai-anthropic-frontier-model-forum/>

introducing autonomous agents to different tasks. Because of this, it is reasonable to expect little coordination between, for example, the UN's efforts to develop rules for LAWs and the work of technical standards bodies to develop AI risk management processes for businesses. Nonetheless, in other areas where coordination would be highly beneficial – like in the development of authoritative AI principles – siloed efforts have led to repetition and fragmentation (Jobin et al., 2019).

The current AI regime complex is not addressing the harms of AI adequately, with initiatives suffering from gridlock, a lack of adoption, or an inability to fulfil their desired purpose. First- and second-order cooperation will pose significant – though not insurmountable – challenges to strengthening global AI governance.

3.1. First-order cooperation problems

First-order cooperation problems stem from international anarchy, with the system lacking a sovereign authority to enforce agreements among state actors. In this context of uncertainty, states perceive themselves to be threatened by other state actors, creating a condition of insecurity. This leads states to increase their capabilities to counter perceived threats (Cerny, 2000), while also undermining cooperation between governments which fear that cooperation will lead to others achieving relative gains (Grieco et al., 1993). The degree to which substantive trust can be built to mitigate these anarchic conditions is contested (Lechner, 2017), yet the formation of international institutions and rules suggests some normative foundations can be established (Bull, 1977).

AI is a policy area particularly subject to first-order cooperation problems. Western European states see in AI a means for comparative economic advantages (Ossewaarde & Gulenc, 2020), while China has singled out AI as a technology that can facilitate an economic “leapfrogging” of competitors (Roberts, Cowls, Morley, et al., 2021). Because AI is a dual-use technology, technical breakthroughs can also benefit national defence. This is a policy aspiration of China, which seeks to use dual-use technologies to promote military-civil fusion through technology transfer between the sectors (Kania, 2019). Fears of China gaining advantages from AI have led the US to situate itself as the leader of “ideologically aligned” countries in opposition to China (Hine & Floridi, 2022). This includes developing defence capabilities, with the US's National Security Commission on AI stressing that the “coalition of democracies” must translate the benefits of AI into a military advantage (Johnson, 2023).

The perceived central role of AI in economic and security competition has led states to enact policies designed to strengthen their international position. The US has taken the most aggressive position by introducing export controls for semiconductors that seek to hinder China's AI development (Schulz, 2022), while also promoting domestic semiconductor production (The White House, 2023). Domestically, China has used industrial policies, such as Made in China 2025, and the political reorganisation of government ministries to promote competitiveness and self-reliance (Beyrer, 2023; Li, 2018). Internationally, it has taken actions to become a norm-maker rather than a norm-taker (Cheng & Zeng, 2023), most notably through exporting technologies as part of its Belt and Road project (Erie & Streinz, 2021). Finally, the EU has been pursuing a policy of “digital sovereignty” designed to strengthen domestic hardware and software capacities, while lessening reliance on foreign technologies (Roberts, Cowls, Casolari, et al., 2021). While arguably incidental, the Brussels Effect associated with EU regulatory efforts gives it international influence

in technology competition by shaping the rules that companies follow internationally (Roberts, Hine, et al., 2023).

Policies designed to improve national competitiveness are not necessarily detrimental to international cooperation. Yet, the perception of AI's centrality to achieving a competitive advantage and narratives of an "arms race" have amplified first-order cooperation problems by undermining mutual trust. For example, a Chinese delegate stated at the first UN Security Council meeting focused on AI that US sanctions undermined prospects for international cooperation on global AI governance.¹⁰ Industry actors have utilised fears of losing an "arms race" to push for significant research and development funding, while simultaneously pushing back against domestic regulatory initiatives (Espinoza & Bounds, 2023; Schmidt, 2020).¹¹ This heightens the risk of unsafe domestic development and deployment of technologies (Scharre, 2021), exacerbating the perception of an arms race and eroding international trust and threatening cooperation.

3.2. *Second-order cooperation problems*

Another challenge to contemporary global AI governance stems from second-order cooperation problems that are associated with institutional dysfunction (Hale et al., 2013). The proliferation of institutions following the Second World War, combined with breakthroughs in transportation and information technology, facilitated greater integration of states and cooperation amongst them (Krasner, 1999; Oneal & Russett, 1997). In turn, many policy areas that were previously the remit of the state fell under the jurisdiction of international regimes. However, success in institutionalisation has complicated international cooperation. Decolonisation, facilitated by institutionalisation, has resulted in more participants in global governance trying to find solutions to problems. Institutional inertia has prevented adaptation to this new reality, while increasing globalised connectivity requires institutions to address more complex problems (Hale et al., 2013). Consider arms control. Previously an interstate problem, it now has to contend with non-state actor access to advanced offensive cyber capabilities, significantly complicating governance (Kello, 2017). Although new international institutions have emerged to address new policy problems, this has arguably exacerbated institutional fragmentation and mandate overlap, limiting the effectiveness of regimes (Hale et al., 2013).

Second-order cooperation problems impact global AI governance. The capacity to develop and regulate AI is currently highly concentrated (Weber, 2017), indicating that multilateral agreement between China, the EU, and the US may be sufficient for mitigating key risks associated with these technologies. Legitimacy questions aside, there is theoretically significant scope for agreement when compared to more multipolar policy areas, particularly given that these states have all promoted coordinating an international response to AI. However, there is little consensus regarding the policy responses needed, with the EU foregrounding new regulations, the US taking a more laissez-faire approach, and China relying on a hybrid approach that utilises "industry self-discipline" and targeted secondary legislation (Roberts, Ziosi, et al., 2023). At an international level, this leads to disagreements in bodies such as the G7 over what types of international governance

¹⁰ <https://media.un.org/en/asset/k1j/k1ji81po8p>

¹¹ Some AI companies have publicly stated they desire to be regulated, yet they are typically referring to a specific type of hypothetical AI safety risk and lobby against other, more immediate regulatory restrictions ("AI Poses Risks Today," 2023). Take OpenAI, which called for the creation of an international AI body, shortly before threatening to leave the EU over the Bloc's regulatory efforts.

instruments should be developed, even amongst those in a so-called “ideological coalition” (Samson, 2023).

The complexity of AI as a policy area exacerbates difficulties. Even leaving aside open questions about how AI should be defined and whether it should be considered a single field of inquiry, the range of policy issues AI encompasses complicates reaching international agreements. There is little agreement among stakeholders as to which policy problems should be prioritised. For instance, there is disagreement between “longtermist” scholars, who focus on the potential existential threats posed by AI, and those more concerned with the harms already materialising, such as bias (“AI Poses Risks Today,” 2023). This divergence is manifesting at the state level, with the UK placing greater emphasis on long-term safety risks than the EU, which has focused on extant harms (Roberts, 2023). Moreover, specific policy remedies face significant technical and philosophical difficulties. As an example, mitigating harmful bias relies on developing a technical measure for fairness *a priori*, which is challenging given international value pluralism complicating what constitutes fairness in different contexts (Binns, 2018; Roberts, Cowls, et al., 2022).

The fragmented international AI landscape further problematises cooperation problems as it allows countries and companies to follow different policies and standards for AI (Cihon et al., 2020). For instance, China is not a member of the GPAI, nor are any Chinese companies involved in the PAI initiative, indicating that two separate AI governance ecosystems are forming. Even where both China and Western states participate in the same international organisations, mandate overlap weakens the legitimacy of any one regime. For instance, the ISO, IEC, IEEE, and ITU have all been active in AI standards-making initiatives.¹² This enables companies to “forum shop”, thus weakening incentives for international cooperation and potentially establishing a “race to the bottom” dynamic.

3.3. *Cooperation problems in practice*

Cooperation problems have impacted nascent global AI governance efforts and will pose significant challenges to strengthening international cooperation going forward. Global agreement on the UNESCO AI Recommendation and by the G20 on the OECD’s AI Principles could be read as forming a thin normative foundation for further cooperation across different AI policy areas. However, the ambiguity of these high-level agreements accommodates different ideological positions, allowing states to interpret them in line with their political and normative priorities (Mittelstadt, 2019). Take AI fairness, a principle supported by all G20 member states, as applied to facial recognition technology. The implementation of this principle in the EU context involves the proposed banning of these technologies, while in China, ethnic-recognition technologies are permissible in the name of order and social stability (Roberts, Cowls, et al., 2022). Efforts to build on these thin normative foundations have faced challenges. Since UNESCO’s AI Ethics Recommendation was signed in November 2021, fewer than a quarter of signatories have worked with the body to implement the policy tools designed to turn the principles into practice.¹³ Signatories have faced no repercussions because the agreement is non-binding. This indicates that,

¹² <https://aistandardshub.org/ai-standards-search/>

¹³ <https://www.unesco.org/en/articles/artificial-intelligence-unesco-calls-all-governments-implement-global-ethical-framework-without>

even when agreement can be made on interventions to mitigate some harms, they are of little practical use without sufficient political buy-in to enact the voluntary initiatives in practice.

In this respect, the COE's draft Convention on AI holds more promise, as it would be legally binding, with states that ratify the treaty expected to translate the Convention into domestic law. However, expectations should be tempered. How many signatories would ratify the treaty and within what period is unclear. Take the COE's Convention on Cybercrime (2001), arguably one of the greatest successes in international cooperation over digital policy. 68 states have ratified the treaty, which promotes international alignment and cooperation, and over 100 countries' domestic cybercrime laws have been influenced by the text (Clough, 2012).¹⁴ Despite this, the cooperation process has been slow, with only five countries initially ratifying the treaty upon its coming into force in 2004. Given the rapid advances in AI, this type of slow-moving agreement will prove inadequate without complementary measures introducing immediate agreement. More problematically, numerous states that were not involved in drafting the Convention refused to ratify it due to a perceived lack of legitimacy, instead pushing for a more representative UN-led process (Clough, 2012). This indicates that even if the COE process does gain traction, institutional fragmentation could cause pushback against the legitimacy of this treaty from countries not involved in the drafting process, such as China and India. Practically, any convention would only address a subset of harms from AI, suggesting significant breakthroughs would also be needed in other AI policy areas.

Given the difficulties associated with interstate cooperation, international private governance could be considered a more promising avenue. This is particularly true for standards-making, which is an area where all major AI powers are still engaged. AI standards committees are predominantly made up of representatives from AI companies which are incentivised to create regulatory consistency across markets that enables them to scale more effectively through lessening regulatory burdens (Cihon, 2019). While standards-making is an innately politicised space,¹⁵ these bodies have thus far proven resilient to wider geopolitical tensions and have produced international governance rules for AI (von Ingersleben-Seip, 2023). However, given the increasing geopolitical pressures that are being put on standards makers – with the US's National Institute for Standards and Technology (NIST) being tasked with devising strategies to “mitigate the PRC's influence” (NIST 2022) – it is unclear how insulated international standards bodies will remain going forward. More importantly, there are limitations to relying on standards bodies to produce positive international outcomes. Some AI regulations, including the EU's draft AI Act, rely on standards bodies to explicate value-laden provisions. Many such provisions are not in a company's commercial interest (Veale & Borgesius, 2021), leading to a principal moral hazard because these companies have the least to gain from developing stringent protections (Miller & Whitford, 2007). Thus, standards bodies are unlikely to provide normatively desirable outcomes for some aspects of AI governance without reform or complementary intergovernmental work.

¹⁴ <https://www.coe.int/en/web/conventions/full-list?module=signatures-by-treaty&treatynum=185>

¹⁵ It is highly lucrative for companies to get their proposals adopted as international standards and a failure to do so requires products to be redesigned and, in some cases, for licence fees to be paid, making it inherently political (Büthe & Mattli, 2011).

4. Pathways forward

Addressing the global risks of AI requires moving from a weak regime complex to the strongest international governance system possible under current geopolitical and institutional conditions. Two pathways forward are notable: (a) developing a new centralised AI institution or (b) strengthening the coordination, capacities, and legitimacy of existing institutions. A necessary first step for assessing potential regimes is developing normative criteria. We focus our assessment on the political legitimacy of regimes, understood broadly in terms of the acceptability of political power, as this is the essential condition of effective governance. We use three criteria for assessing the political legitimacy of a potential regime. The two substantive criteria that define the political legitimacy of a regime are:

- (1) Its ability to enable states and other actors to coordinate their behaviour in mutually beneficial ways (Buchanan & Keohane, 2006);
- (2) The presence of democratic procedures for decision-making (Erman & Furendal, 2022);

Alongside this, there is a necessary operationalising criterion:

- (3) Its viability considering geopolitical and institutional realities.

In short, the regime should support internationally desirable outcomes, have a high degree of input and/ or procedural legitimacy, and be achievable. We argue that strengthening the regime complex for AI offers a more desirable and realistic path forward based on these criteria.

4.1. *Institution building*

Several public figures have suggested establishing an international AI governance body akin to the IPCC or IAEA. These international bodies are given as examples of expert governance regimes that can be established under heightened political tensions to mitigate global risks (Chowdhury, 2023; Marcus & Reuel, 2023). However, such institutions would only provide a partial solution for coordinating behaviour in a way that supports mutually beneficial outcomes. Climate governance is decentralised, with many institutions involved (Keohane & Victor, 2011), so an IPCC-like body would aid coordination, but it would need to be accompanied by other types of institutional reform to secure desirable outcomes (an option discussed below). In contrast, the IAEA model offers a more centralised solution, with this body at the heart of global nuclear governance. The problem with emulating a centralised model is that nuclear and AI governance are not similar types of policy problems. AI policy is not a clearly defined, with disagreement over what falls into the field and what constitutes harm. It is also decentralised as it does not face the same physical bottlenecks in materials that nuclear does. Finally, it has a cross-cutting impact in terms of the sectors that will be impacted and the range of upstream and downstream harms it can cause. These features suggest that a single AI body is unlikely to be an effective mechanism for coordinating relevant stakeholders.

A new AI body established through a democratic procedure in a multilateral forum, such as the UN, would have a high degree of procedural legitimacy, but would not be unimpeachable (Erman & Furendal, 2022). An IAEA-like body for AI would be at higher risk of regulatory capture than the original IAEA because of the commercial incentives for Big Tech companies to shape governance, potentially undermining a body's democratic mandate. Indeed, these companies are

already attempting to shift the narrative from the harms their products are currently causing to those that are more speculative, which has arguably been successful in the framing the UK's Summit ("AI Poses Risks Today" 2023). Moreover, while centralised regime mandates are often brittle, AI is at particular risk due to the pace of development (Cihon et al., 2020). Updating an organisation's remit is possible, but it may require a new democratic mandate, which would be time-consuming. As such, there is a likely trade-off between the outcome and procedural legitimacy of the organisation.

Finally, there is the question of the viability of creating a meaningfully empowered centralised AI body. The type of international bodies offered as exemplars for AI governance were established through formal diplomatic procedures, under very different geopolitical and institutional circumstances. The IAEA was created in 1957, during a period of proliferation in international institutions (Hale et al., 2013), and only gained substantive powers after the Non-Proliferation Treaty was concluded in 1968 (Clarke, 2017). Likewise, the IPCC emerged from decades of expert international cooperation that provided the foundation for the body's work (De Pryck & Hulme, 2022). Contemporary cooperation problems make establishing formal international institutions more difficult (Grigorescu, 2020). When combined with the nascency of international governance efforts for AI, there will be significant difficulties in emulating these models. Furthermore, the success of institution building and ongoing agreement in nuclear arms control was founded on a fear of existential risk, something that is not agreed upon for AI, weakening incentives for cooperation ("AI Poses Risks Today" 2023). This is not to say that establishing such a body is impossible under contemporary geopolitical conditions, particularly if a global shock acts as a catalyst for action, but reaching agreement on a strong institution would be difficult and time-consuming.

4.2. *Strengthening the regime complex*

An alternative to developing a centralised international AI body is strengthening the weak regime complex. This could take many forms, such as improving information sharing, developing joint partnerships, and creating conflict resolution mechanisms between nodes (Galaz et al., 2012). These strategies would lessen fragmentation, improving the authoritativeness of governance initiatives. Strengthening the regime complex could also involve developing new institutions to fill governance gaps. However, due to the risk of further fragmentation and dilution of authority, this can be seen as a secondary priority to improving coordination, capacities, and legitimacy of what already exists.

Strengthening the AI regime complex would allow for the advantages of a decentralised approach to materialise. For instance, even if first-order competition prohibits cooperation in some policy areas, the polycentric nature of the model would allow for cooperation in others (Cihon et al., 2020). A strong regime complex would also have the flexibility to adapt to new policy challenges, and clarifying institutional remits and strengthening communication channels would help define which nodes(s) should address problems provoked by new AI developments. Perhaps most importantly, the "polycentric ordering" of the regime complex means no external authority can impose coordination, as it is only generated by committed participants (Abbott, 2014). While this could be viewed negatively, decentralised action from myriad stakeholders at different levels of governance can encourage mutually reinforcing commitments, facilitating catalytic change (Keohane & Victor, 2011). This means that even if geopolitical conditions limit substantive

interstate cooperation and action, other stakeholders can still create positive change through mutually reinforcing actions (Abbott, 2014).

There are potential drawbacks to a regime complex model. Notably, Big Tech companies may more easily ignore a decentralised international structure, or have certain parts captured due to the greater resource constraints civil society organisations will face in being represented at the various fora (Cihon et al., 2020). Alongside this, some goals may fall through the cracks of the regime complex if no institution considers them within their remit (Alter & Raustiala, 2017). However, these are not inherent features of the system design, but problems with the coordination and authority of institutions, which can be addressed. The impact of regime capture is also lesser in a regime complex than a centralised regime, due to distributed impacts. Furthermore, an organisation with a fixed mandate is more likely to miss addressing some problems than a decentralised model.

The proposal to strengthen the current regime complex may sound fanciful given current fragmentation and the characteristics of AI as a policy area that lends itself to fragmented and overlapping governance. Yet, there is precedent in other areas of international policymaking, notably climate. Climate governance, like AI, is not a single policy problem and instead involves different issues like biodiversity and carbon emissions. After decades of failure to reach meaningful global agreements, capped by the unfruitful 2009 Copenhagen Summit, the focus on climate governance moved from developing a centralised regime to leveraging the polycentric order. This includes international regimes like the IPCC, the UNFCCC, and bilateral initiatives (Keohane & Victor, 2011). Notably, the climate regime complex is also multi-level, with local governments and private companies encouraged to enact innovative policies suited to their specific challenges and responsibilities (Jänicke, 2017). This is a new and largely untested model of governance (Jordan et al., 2015), with some scepticism due, on account of evidence of both governments and companies reneging on their climate pledges (Tollefson, 2022). Nonetheless, existing research has highlighted some promising indicators, like the boldest climate pledges being the most credible (Victor et al., 2022). More importantly, although this model is flawed, compared with other policy areas and environmental policy outcomes from periods where efforts were made to establish a centralised regime, there have been remarkable successes (Abbott, 2014; Hale, 2017).

On the surface, a centralised AI body could be perceived as more legitimate than a decentralised regime because of its democratic mandate from states. However, when considering democratic input beyond the state, a decentralised model facilitates consideration of more issues and stakeholders. This does come with the aforementioned risk of resource constraints for civil society actors (Cihon et al., 2020), but strong coordination can help mitigate this. Moreover, a polycentric model allows for input at different levels of governance, not just international organisations. This is particularly important given the legitimacy problem affecting international organisations in terms of citizen support (Dellmuth et al., 2022).

In terms of viability, strengthening regime complex coordination requires action from more actors than in a centralised regime and international agreement on responsibility allocation. This will be challenging considering the cross-cutting nature of AI and because some actors benefit from fragmented institutions. Nonetheless, coordination is possible, and there are numerous examples of international bodies promoting coordination through mechanisms like memoranda of understanding and joint projects. For instance, while the OECD and UNESCO have largely

been fragmented in their AI policy work, there is a precedent of collaboration between the OECD and UN bodies, such as the joint Tax Inspectors Without Borders initiative. It is possible to imagine a similar model of collaboration working for AI, including in areas like AI auditing.

Given the non-binding nature of the model, a plurality of actors would need to act as “norm entrepreneurs” in developing mechanisms or making unilateral commitments (Finnemore & Sikkink, 1998). There are promising signs that states and other stakeholders would be willing to undertake this. Norm entrepreneurs are undertaking unilateral actions with an international impact, including the EU’s AI regulations that will have an international impact through the so-called “Brussels Effect” (Bradford, 2020; Siegmann & Anderljung, 2022), and recent US and Chinese efforts inviting voluntary pledges from the companies that are developing cutting-edge AI systems, and the UK’s AI Summit. Beyond state-level initiatives, US subnational governments have been introducing AI regulations, and Norway’s Sovereign Wealth Fund recently announced its intention to establish investor standards for assessing AI and called for clearer government regulation to base this upon (Milne & Martin, 2023). If sufficiently aligned, this type of action can be mutually reinforcing in addressing various AI harms from different levels of governance.

Finally, the value-laden nature of AI could prove challenging for a regime complex model that relies on unilateral action. The regime complex for climate has succeeded in achieving pledges from a range of government and non-government entities to cut emissions (Hale, 2017). While there are different ways of measuring this type of pledge, it is arguably simpler than an innately value-laden equivalent in the field of AI governance, such as reducing harmful bias (Binns, 2018). This complexity could make decentralised action more challenging and undermine accountability. Nonetheless, expert work to develop authoritative benchmarks could aid peer pressure and mutually reinforcing change.

5. Recommendations

The above discussion shows that strengthening the regime complex for AI is a desirable path forward, but one that will require work to avoid a continuation of current fragmentation. For this reason, we offer recommendations designed to support the transformation from a weak regime complex to one more capable of addressing the myriad challenges posed by AI. We align these recommendations with our three criteria for political legitimacy.

5.1. Enabling coordination

To enable states to coordinate within the regime complex, it is important to strengthen ties between nodes in the network. This needs to happen at two levels: first, between key international bodies whose remits related to AI overlap, such as UNESCO and the OECD. Forthcoming work by the UN High-Level Advisory Body to map the international AI ecosystem and provide recommendations is an important opportunity to begin addressing this overlap. This could be followed by negotiations to agree on remits, as is common when international organisations have overlapping jurisdiction. Second, because of the polycentric order that involves state and non-state actors, formal mechanisms for strengthening information sharing and partnerships will not be applicable in many cases. Instead, focusing efforts on aligning nodes in the network around common goals would be beneficial for supporting self-reinforcing action from a variety of actors. Of course, given AI’s ambiguous, value-laden, and cross-cutting nature, this type of alignment is difficult, but not impossible.

A new expert UN body could be desirable for fulfilling this specific expert role within the regime complex. Yet, scepticism is warranted both in terms of whether it could be established and, if established, whether it would have meaningful authority. In its absence, it is important for authoritative expert guidance to be developed for stakeholders. The OECD's nature and existing role in international cooperation on AI make it a viable choice for this role.

The OECD is an institution relatively insulated from geopolitics due to its technocratic nature. It lacks strong carrots and sticks to enforce compliance, instead using soft governance mechanisms to push for cooperation and alignment. It is a “meditative” forum for constructing and disseminating research and policy ideas that subsequently shape policy choices and implementation at a national level, frequently including hard regulations (Mahon & McBride, 2009). In this sense, the OECD influences policy through “normative governance”, which promotes policy alignment by challenging and changing the approaches of national governments (Woodward, 2009). Moreover, the OECD supports peer pressure among states by developing indicators, standard forms of measurement, and rankings that permit cross-country comparisons (Klasen, 2017).

Building on its existing network of experts, the OECD could produce expert policy documents that could act as a foundation for greater alignment, such as in-depth comparisons of how countries are managing high-risk AI systems, an incident database of misuse, and economic data about which states are benefiting or being harmed by AI developments. A more aspirational policy product could be an IPCC-like report, produced by global experts, which outlines consensus and outstanding questions on policy problems and solutions. Like with IPCC reports, summaries could then be sent to state representatives for approval, providing scientific and political credibility to the findings (Hale, 2017). This could risk politicisation, yet it could also clarify where expert and international alignment lies, facilitating policy agreement.

Two risks are associated with relying on the OECD as an international expert authority for AI. First, the OECD's work could be perceived as illegitimate because it is a predominantly European organisation, undermining the authoritativeness of its outputs. Having work delegated to the OECD from a more authoritative organisation could enhance its legitimacy. There is a strong precedent of the OECD playing this “palliative governance” role by using its technical expertise to grease the “wheels of global governance” in support of other bodies (Woodward, 2009). One example is the OECD's work on taxation; the G20 delegated the development of an international framework designed to mitigate the exploitation of tax gaps, which 142 countries agreed to (Fung, 2017).¹⁶ While scepticism over G20 agreement to delegate work of this sort is warranted, it is not unreasonable given the high-profile political attention AI has received, and it is arguably more likely than creating an effective new institution. In the meantime, informal support for the OECD's work from other international organisations, combined with continued efforts to include a range of stakeholders in expert discussions¹⁷ and to undertake “enhanced engagement” with potentially sceptical non-members such as China, would help to mitigate risks. Second, the OECD may not be sufficiently resourced to achieve the expert consensus needed. Because of this, it could be beneficial to rely on an authoritative and independent, AI-specific body

¹⁶ <https://www.oecd.org/tax/beps/inclusive-framework-on-beps-composition.pdf>

¹⁷ <https://oecd.ai/en/community?role=&workingGroupId=&terms=&countryId=&stakeholderTypeId=&page=1>

in the longer term, but given current coordination problems, focus should be on an intermediary solution, which the OECD can provide.

5.2. *Enhancing legitimacy*

Ensuring that different actors are contributing appropriately is equally important for developing an effective regime complex. Institutions in regime complexes differ in how democratic their procedures are. For highly value-laden AI governance work, institutions must be inclusive and accountable. Otherwise, the legitimacy of the work can be questioned, and outcomes may be undesirable (Keohane & Victor, 2011). Mitigating this problem will foremost require addressing the problems in the appropriate bodies.

Current or proposed work on highly value-laden issues in standards bodies is one example of where caution should be taken regarding input and procedural legitimacy. Because industry actors dominate standards bodies, suggestions have been made to increase civil society participation (Galvagna, 2023).¹⁸ However, existing research has shown a significant negative association between the technical complexity of regulatory proposals and the degree of “mobilised dissent” by civil society actors (Pagliari & Young, 2016). This indicates that civil society actors, who have an interest in the outcomes of AI standards-making, are unlikely to be able to take part meaningfully due to resource and expertise limitations. On top of this, enhanced civil society participation risks exacerbating politicisation in standards bodies, particularly in value-laden areas like human rights, where substantive conceptual differences are present between the US and China (Angle, 2002).

In short, there appears to be a trade-off between input legitimacy in standards bodies and the risk of further politicisation undermining good governance. Accordingly, moving the level of abstraction from focusing on the processes of specific bodies to questioning whether they are appropriate fora for the aspect of AI governance would prove a fruitful avenue for strengthening the procedural legitimacy of the regime complex.

5.3. *Improving viability*

Improving the viability of a strong regime complex for AI first requires shifting the discussion, which thus far has focused on developing a new AI body. It took decades of cooperation failure in climate governance to move towards a more decentralised model. We should not make the same mistake for AI. The UN High-Level Advisory Body on AI should be aware of this fact and use it to push the conversation towards strengthening the existing regime complex. A strong regime complex for AI will be imperfect, but it will make possible immediate change. Given the pace of AI development, it is essential not to wait for a new AI body that may never exist in a form capable of tackling the risk associated with AI.

Reference list

10 Downing Street. (2023). *UK to host first global summit on Artificial Intelligence*. GOV.UK.
<https://www.gov.uk/government/news/uk-to-host-first-global-summit-on-artificial-intelligence>

¹⁸ https://g7digital-tech-2023.go.jp/topics/pdf/pdf_20230430/annex5.pdf

- Abbott, K. W. (2014). Strengthening the Transnational Regime Complex for Climate Change. *Transnational Environmental Law*, 3(1), 57–88.
<https://doi.org/10.1017/S2047102513000502>
- Alter, K. J., & Raustiala, K. (2017). *The Rise of International Regime Complexity* (SSRN Scholarly Paper 3085043). <https://papers.ssrn.com/abstract=3085043>
- Altman, S., Brockman, G., & Sutskever, I. (2023). *Governance of superintelligence*. OpenAI.
<https://openai.com/blog/governance-of-superintelligence>
- Angle, S. (2002). *Human Rights in Chinese Thought: A Cross-Cultural Inquiry*. Cambridge University Press.
- Bak-Coleman, J., Bergstrom, C. T., Jacquet, J., Mickens, J., Tufekci, Z., & Roberts, T. (2023). Create an IPCC-like body to harness benefits and combat harms of digital tech. *Nature*, 617(7961), 462–464. <https://doi.org/10.1038/d41586-023-01606-9>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.
<https://doi.org/10.1145/3442188.3445922>
- Beyrer, P. (2023). *Beijing's Tech Restructuring Signals Enhanced Competition With the United States*. Asia Society. <https://asiasociety.org/policy-institute/beijings-tech-restructuring-signals-enhanced-competition-united-states>
- Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 149–159.
<https://proceedings.mlr.press/v81/binns18a.html>
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv.
<https://doi.org/10.48550/arXiv.2108.07258>
- Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press.
- Buchanan, A., & Keohane, R. O. (2006). The Legitimacy of Global Governance Institutions. *Ethics & International Affairs*, 20(4), 405–437. <https://doi.org/10.1111/j.1747-7093.2006.00043.x>
- Bull, H. (1977). *The Anarchical Society*. Macmillan Education UK. <https://doi.org/10.1007/978-1-349-24028-9>
- Büthe, T., & Mattli, W. (2011). The New Global Rulers: The Privatization of Regulation in the World Economy. In *The New Global Rulers*. Princeton University Press.
<https://doi.org/10.1515/9781400838790>
- Cerny, P. G. (2000). The New Security Dilemma: Divisibility, Defection and Disorder in the Global Era. *Review of International Studies*, 26(4), 623–646.
- Cheng, J., & Zeng, J. (2023). Shaping AI's Future? China in Global AI Governance. *Journal of Contemporary China*, 32(143), 794–810. <https://doi.org/10.1080/10670564.2022.2107391>
- Chowdhury, R. (2023, April 6). AI Desperately Needs Global Oversight. *Wired*.
<https://www.wired.com/story/ai-desperately-needs-global-oversight/>
- Cihon, P. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research & Development*. Future of Humanity Institute. https://www.fhi.ox.ac.uk/wp-content/uploads/Standards_FHI-Technical-Report.pdf
- Cihon, P., Maas, M. M., & Kemp, L. (2020). Fragmentation and the Future: Investigating Architectures for International AI Governance. *Global Policy*, 11(5), 545–556.
<https://doi.org/10.1111/1758-5899.12890>
- Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. *Information*, 12(7), Article 7. <https://doi.org/10.3390/info12070275>

- Clarke, M. (2017). Weapons of Mass Destruction: Incremental Steps. In T. Hale & D. Held, *Beyond Gridlock*. John Wiley & Sons.
- Clough, J. (2012). The Council of Europe Convention on Cybercrime: Defining 'Crime' in a Digital World. *Criminal Law Forum*, 23(4), 363–391. <https://doi.org/10.1007/s10609-012-9183-3>
- Council of Europe creates rules for Artificial Intelligence. (2023). AlgorithmWatch. <https://algorithmwatch.org/en/artificial-intelligence-council-of-europe/>
- De Pryck, K., & Hulme, M. (Eds.). (2022). *A Critical Assessment of the Intergovernmental Panel on Climate Change*. Cambridge University Press. <https://doi.org/10.1017/9781009082099>
- Dellmuth, L., Scholte, J. A., Tallberg, J., & Verhaegen, S. (2022). The Elite–Citizen Gap in International Organization Legitimacy. *American Political Science Review*, 116(1), 283–300. <https://doi.org/10.1017/S0003055421000824>
- Drexel, B., & Depp, M. (2023, June 13). Every Country Is on Its Own on AI. *Foreign Policy*. <https://foreignpolicy.com/2023/06/13/ai-regulation-international-nuclear/>
- Erie, M. S., & Streinz, T. (2021). *The Beijing Effect: China's "Digital Silk Road" as Transnational Data Governance* (SSRN Scholarly Paper ID 3810256). Social Science Research Network. <https://papers.ssrn.com/abstract=3810256>
- Erman, E., & Furendal, M. (2022). Artificial Intelligence and the Political Legitimacy of Global Governance. *Political Studies*, 00323217221126665. <https://doi.org/10.1177/00323217221126665>
- Espinoza, J., & Bounds, A. (2023, May 7). Tech groups call for changes to EU data-sharing proposals. *Financial Times*. <https://www.ft.com/content/8ce1f3a9-1859-499f-972e-d4801b9d44d1>
- EU-US Trade and Technology Council. (n.d.). European Commission. Retrieved April 23, 2023, from https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/stronger-europe-world/eu-us-trade-and-technology-council_en
- Fatima, S., Dawson, G. S., Desouza, K. C., & Denford, J. S. (2021). *Winners and losers in the fulfillment of national artificial intelligence aspirations*. Brookings. <https://www.brookings.edu/blog/techtank/2021/10/21/winners-and-losers-in-the-fulfilment-of-national-artificial-intelligence-aspirations/>
- Finkelstein, L. S. (1995). What Is Global Governance? *Global Governance*, 1(3), 367–372.
- Finnemore, M., & Sikkink, K. (1998). International Norm Dynamics and Political Change. *International Organization*, 52(4), 887–917. <https://doi.org/10.1162/002081898550789>
- Fung, S. (2017). *The Questionable Legitimacy of the OECD/G20 BEPS Project* (SSRN Scholarly Paper 3087256). <https://papers.ssrn.com/abstract=3087256>
- G7 Hiroshima Leaders' Communiqué. (2023). G7 Hiroshima Summit. <https://www.g7hiroshima.go.jp/en/documents/>
- G20 Osaka Leaders' Declaration. (2019). G20 Osaka Summit 2019. https://mofa.go.jp/policy/economy/g20_summit/osaka19/documents/final_g20_osaka_leaders_declaration.html
- Galaz, V., Crona, B., Österblom, H., Olsson, P., & Folke, C. (2012). Polycentric systems and interacting planetary boundaries—Emerging governance of climate change—ocean acidification—marine biodiversity. *Ecological Economics*, 81, 21–32. <https://doi.org/10.1016/j.ecolecon.2011.11.012>
- Galvagna, C. (2023). *Discussion paper: Inclusive AI governance*. Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/report/inclusive-ai-governance/>
- GPAI Terms of Reference. (2020). Global Partnership on AI. <https://gpai.ai/about/gpai-terms-of-reference.pdf>
- Grieco, J., Powell, R., & Snidal, D. (1993). The Relative-Gains Problem for International Cooperation. *The American Political Science Review*, 87(3), 727–743. <https://doi.org/10.2307/2938747>

- Grigorescu, A. (2020). *The Ebb and Flow of Global Governance: Intergovernmentalism versus Nongovernmentalism in World Politics*. Cambridge University Press.
<https://doi.org/10.1017/9781108850049>
- Hale, T. (2017). Climate Change: From Gridlock to Catalyst. In D. Held & T. Hale, *Beyond Gridlock*. John Wiley & Sons.
- Hale, T., & Held, D. (2017). *Beyond Gridlock*. Polity Press.
<http://ebookcentral.proquest.com/lib/oxford/detail.action?docID=5077165>
- Hale, T., Held, D., & Young, K. (2013). *Gridlock: Why Global Cooperation is Failing when We Need It Most*. Polity.
- Hazell, J. (2023). *Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns* (arXiv:2305.06972). arXiv. <https://doi.org/10.48550/arXiv.2305.06972>
- Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, 12(1).
<https://policyreview.info/essay/chatgpt-and-ai-act>
- Hine, E., & Floridi, L. (2022). Artificial intelligence with American values and Chinese characteristics: A comparative analysis of American and Chinese governmental AI policies. *AI & SOCIETY*. <https://doi.org/10.1007/s00146-022-01499-8>
- Ho, L., Barnhart, J., Trager, R., Bengio, Y., Brundage, M., Carnegie, A., Chowdhury, R., Dafoe, A., Hadfield, G., Levi, M., & Snidal, D. (2023). *International Institutions for Advanced AI* (arXiv:2307.04699). arXiv. <http://arxiv.org/abs/2307.04699>
- Jänicke, M. (2017). The Multi-level System of Global Climate Governance – the Model and its Current State. *Environmental Policy and Governance*, 27(2), 108–121.
<https://doi.org/10.1002/eet.1747>
- Jelinek, T., Wallach, W., & Kerimi, D. (2021). Policy brief: The creation of a G20 coordinating committee for the governance of artificial intelligence. *AI and Ethics*, 1(2), 141–150.
<https://doi.org/10.1007/s43681-020-00019-y>
- Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389–399. <https://doi.org/10.1038/s42256-019-0088-2>
- Johnson, J. (2023). AI-security dilemma: Insecurity, mistrust, and misperception under the nuclear shadow. In J. Johnson (Ed.), *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (p. 0). Oxford University Press.
<https://doi.org/10.1093/oso/9780192858184.003.0005>
- Jordan, A. J., Huitema, D., Hildén, M., van Asselt, H., Rayner, T. J., Schoenefeld, J. J., Tosun, J., Forster, J., & Boasson, E. L. (2015). Emergence of polycentric climate governance and its future prospects. *Nature Climate Change*, 5(11), Article 11.
<https://doi.org/10.1038/nclimate2725>
- Kak, A., Myers West, S., Hanna, A., Gebru, T., Gahntz, M., Talat, Z., & Khan, M. (2023, April 13). General Purpose AI Poses Serious Risks, Should Not Be Excluded From the EU's AI Act. *AI Now Institute*. <https://ainowinstitute.org/publication/gpai-is-high-risk-should-not-be-excluded-from-eu-ai-act>
- Kania, E. B. (2019). Chinese Military Innovation in the AI Revolution. *The RUSI Journal*, 164(5–6), 26–34. <https://doi.org/10.1080/03071847.2019.1693803>
- Kello, L. (2017). Cybersecurity: Gridlock and Innovation. In T. Hale & D. Held (Eds.), *Beyond Gridlock*. John Wiley & Sons.
- Keohane, R. O., & Victor, D. G. (2011). The Regime Complex for Climate Change. *Perspectives on Politics*, 9(1), 7–23. <https://doi.org/10.1017/S1537592710004068>
- Klasen, A. (2017). Trade: Gridlock and Resilience. In T. Hale & D. Held, *Beyond Gridlock*. John Wiley & Sons.
- Krasner, S. D. (1999). *Sovereignty: Organized Hypocrisy*. Princeton University Press.
<https://www.jstor.org/stable/j.ctt7s9d5>
- Küspert, S., Moës, N., & Dunlop, C. (2023). *The value chain of general-purpose AI*. Ada Lovelace Institute. <https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/>

- Lechner, S. (2017, November 20). *Anarchy in International Relations*. Oxford Research Encyclopedia of International Studies. <https://doi.org/10.1093/acrefore/9780190846626.013.79>
- Li, L. (2018). China's manufacturing locus in 2025: With a comparison of "Made-in-China 2025" and "Industry 4.0." *Technological Forecasting and Social Change*, 135, 66–74. <https://doi.org/10.1016/j.techfore.2017.05.028>
- Maas, M. M. (2019). How viable is international arms control for military artificial intelligence? Three lessons from nuclear weapons. *Contemporary Security Policy*, 40(3), 285–311. <https://doi.org/10.1080/13523260.2019.1576464>
- Mahon, R., & McBride, S. (2009). Standardizing and disseminating knowledge: The role of the OECD in global governance. *European Political Science Review*, 1(1), 83–101. <https://doi.org/10.1017/S1755773909000058>
- Malik, A. (2023, March 14). Duolingo launches new subscription tier with access to AI tutor powered by GPT-4. *TechCrunch*. <https://techcrunch.com/2023/03/14/duolingo-launches-new-subscription-tier-with-access-to-ai-tutor-powered-by-gpt-4/>
- Marcus, G., & Reuel, A. (2023, April 18). The world needs an international agency for artificial intelligence, say two AI experts. *The Economist*. <https://www.economist.com/by-invitation/2023/04/18/the-world-needs-an-international-agency-for-artificial-intelligence-say-two-ai-experts>
- McGarr, T. (2023, April 19). *Generative AI and Standards: An overview*. TechUK. <https://www.techuk.org/resource/aiweek2023-bsi-wed.html>
- Miller, G. J., & Whitford, A. B. (2007). The Principal's Moral Hazard: Constraints on the Use of Incentives in Hierarchy. *Journal of Public Administration Research and Theory*, 17(2), 213–233. <https://doi.org/10.1093/jopart/mul004>
- Milne, R., & Martin, K. (2023, April 28). Norway's \$1.4tn wealth fund calls for state regulation of AI. *Financial Times*. <https://www.ft.com/content/594a4f52-eb98-4da2-beca-4addcf9777c4>
- Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), Article 11. <https://doi.org/10.1038/s42256-019-0114-4>
- Multistakeholder Experts Group Report*. (2022). Global Partnership on AI. <https://gpai.ai/projects/gpai-multistakeholder-expert-group-report-november-2022.pdf>
- National Institute of Standards and Technology. (2022). *A Report to Congress: Steps to Implement Recommendations Regarding "U.S. Leadership in Artificial Intelligence (AI): A Plan for Federal Engagement in Developing Technical Standards and Related Tools"* (p. 12). https://www.nist.gov/system/files/documents/2022/08/31/21-H-506-NIST-U_S_Leadership_in_AI_Report_to_Congress_Report.pdf
- NATO. (2021). *Summary of the NATO Artificial Intelligence Strategy*. NATO. https://www.nato.int/cps/en/natohq/official_texts_187617.htm
- Oneal, J. R., & Russett, B. M. (1997). The Classical Liberals Were Right: Democracy, Interdependence, and Conflict, 1950–1985. *International Studies Quarterly*, 41(2), 267–293. <https://doi.org/10.1111/1468-2478.00042>
- Ossewaarde, M., & Gulenc, E. (2020). National Varieties of Artificial Intelligence Discourses: Myth, Utopianism, and Solutionism in West European Policy Expectations. *Computer*, 53(11), 53–61. <https://doi.org/10.1109/MC.2020.2992290>
- Pagliari, S., & Young, K. (2016). The interest ecology of financial regulation: Interest group plurality in the design of financial regulatory policies. *Socio-Economic Review*, 14(2), 309–337. <https://doi.org/10.1093/ser/mwv024>
- PAI Staff. (2022, May 25). *Recommendations in Practice: Meta Researchers Apply PAI Guidance*. Partnership on AI. <https://partnershiponai.org/resource/recommendations-in-practice-meta-researchers-apply-pai-guidance/>

- Pause Giant AI Experiments: An Open Letter. (2023). *Future of Life Institute*.
<https://futureoflife.org/open-letter/pause-giant-ai-experiments/>
- Roberts, H. (2023, June 26). *AI in the EU and UK: Two approaches to regulation and international leadership*. UK in a Changing Europe. <https://ukandeu.ac.uk/ai-in-the-eu-and-uk-two-approaches-to-regulation-and-international-leadership/>
- Roberts, H., Babuta, A., Morley, J., Thomas, C., Taddeo, M., & Floridi, L. (2022). *Artificial Intelligence Regulation in the United Kingdom: A Path to Global Leadership?* (SSRN Scholarly Paper 4209504). <https://doi.org/10.2139/ssrn.4209504>
- Roberts, H., Cowls, J., Casolari, F., Morley, J., Taddeo, M., & Floridi, L. (2021). Safeguarding European values with digital sovereignty: An analysis of statements and policies. *Internet Policy Review*, 10(3). <https://policyreview.info/articles/analysis/safeguarding-european-values-digital-sovereignty-analysis-statements-and-policies>
- Roberts, H., Cowls, J., Hine, E., Mazzi, F., Tsamados, A., Taddeo, M., & Floridi, L. (2021). Achieving a ‘Good AI Society’: Comparing the Aims and Progress of the EU and the US. *Science and Engineering Ethics*, 27(6), 68. <https://doi.org/10.1007/s11948-021-00340-7>
- Roberts, H., Cowls, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2022). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. *The Information Society*, 0(0), 1–19. <https://doi.org/10.1080/01972243.2022.2124565>
- Roberts, H., Cowls, J., Morley, J., Taddeo, M., Wang, V., & Floridi, L. (2021). The Chinese approach to artificial intelligence: An analysis of policy, ethics, and regulation. *AI & SOCIETY*, 36(1), 59–77. <https://doi.org/10.1007/s00146-020-00992-2>
- Roberts, H., Hine, E., & Floridi, L. (2023). *Digital Sovereignty, Digital Expansionism, and the Prospects for Global AI Governance* (SSRN Scholarly Paper 4483271). <https://papers.ssrn.com/abstract=4483271>
- Roberts, H., Ziosi, M., & Cailean, O. (2023). *A Comparative Framework for AI Regulatory Policy*. CEIMIA. <https://ceimia.org/wp-content/uploads/2023/02/Comparative-Framework-for-AI-Regulatory-Policy.pdf>
- Roger, C. B. (2020). The Move to Informality. In C. B. Roger (Ed.), *The Origins of Informality: Why the Legal Foundations of Global Governance are Shifting, and Why It Matters* (p. 0). Oxford University Press. <https://doi.org/10.1093/oso/9780190947965.003.0001>
- Samson, P. (2023). *On Advancing Global AI Governance*. Centre for International Governance Innovation. <https://www.cigionline.org/articles/on-advancing-global-ai-governance/>
- Scharre, P. (2021, June 28). *Debunking the AI Arms Race Theory*. Texas National Security Review. <https://tnsr.org/2021/06/debunking-the-ai-arms-race-theory/>
- Scharre, P., & Lamberth, M. (2022). *Artificial Intelligence and Arms Control* (arXiv:2211.00065). arXiv. <https://doi.org/10.48550/arXiv.2211.00065>
- Schmidt, E. (2020, February 27). Opinion | Eric Schmidt: I Used to Run Google. Silicon Valley Could Lose to China. *The New York Times*. <https://www.nytimes.com/2020/02/27/opinion/eric-schmidt-ai-china.html>
- Schmitt, L. (2022). Mapping global AI governance: A nascent regime in a fragmented landscape. *AI and Ethics*, 2(2), 303–314. <https://doi.org/10.1007/s43681-021-00083-y>
- Schulz, M. (2022, September 7). Passage of the CHIPS and Science Act: What does this mean for U.S. export controls? *Reuters*. <https://www.reuters.com/legal/legalindustry/passage-chips-science-act-what-does-this-mean-us-export-controls-2022-09-07/>
- Siegmann, C., & Anderljung, M. (2022). *The Brussels Effect and Artificial Intelligence*. Centre for the Governance of AI. <https://www.governance.ai/research-paper/brussels-effect-ai>
- Stanley-Lockman, Z., & Christie, E. H. (2021, October 25). *NATO Review—An Artificial Intelligence Strategy for NATO*. NATO Review. <https://www.nato.int/docu/review/articles/2021/10/25/an-artificial-intelligence-strategy-for-nato/index.html>

- Stop talking about tomorrow's AI doomsday when AI poses risks today. (2023). *Nature*, 618(7967), 885–886. <https://doi.org/10.1038/d41586-023-02094-7>
- Tallberg, J., Erman, E., Furendal, M., Geith, J., Klamberg, M., & Lundgren, M. (2023). *The Global Governance of Artificial Intelligence: Next Steps for Empirical and Normative Research* (arXiv:2305.11528). arXiv. <http://arxiv.org/abs/2305.11528>
- The OECD AI Principles*. (2019). <https://oecd.ai/en/ai-principles>
- The White House. (2023, April 27). *Remarks by National Security Advisor Jake Sullivan on Renewing American Economic Leadership at the Brookings Institution*. The White House. <https://www.whitehouse.gov/briefing-room/speeches-remarks/2023/04/27/remarks-by-national-security-advisor-jake-sullivan-on-renewing-american-economic-leadership-at-the-brookings-institution/>
- Tollefson, J. (2022). Climate pledges from top companies crumble under scrutiny. *Nature*. <https://doi.org/10.1038/d41586-022-00366-2>
- UN Press. (2023, July 18). *Secretary-General Urges Security Council to Ensure Transparency, Accountability, Oversight, in First Debate on Artificial Intelligence*. <https://press.un.org/en/2023/sgsm21880.doc.htm>
- UNESCO adopts first global standard on the ethics of artificial intelligence. (2022). UNESCO. <https://www.unesco.org/en/articles/unesco-adopts-first-global-standard-ethics-artificial-intelligence>
- Veale, M., & Borgesius, F. Z. (2021). *Demystifying the Draft EU Artificial Intelligence Act*. SocArXiv. <https://doi.org/10.31235/osf.io/38p5f>
- Veale, M., Matus, K., & Gorwa, R. (2023). *AI and Global Governance: Modalities, Rationales, Tensions*. SocArXiv. <https://doi.org/10.31235/osf.io/ubxgk>
- Verma, P., & Oremus, W. (2023, April 14). ChatGPT invented a sexual harassment scandal and named a real law prof as the accused. *Washington Post*. <https://www.washingtonpost.com/technology/2023/04/05/chatgpt-lies/>
- Victor, D. G., Lumkowsky, M., & Dannenberg, A. (2022). Determining the credibility of commitments in international climate policy. *Nature Climate Change*, 12(9), Article 9. <https://doi.org/10.1038/s41558-022-01454-x>
- von Ingersleben-Seip, N. (2023). Competition and cooperation in artificial intelligence standard setting: Explaining emergent patterns. *Review of Policy Research*, 00, 1–30. <https://doi.org/10.1111/ropr.12538>
- Weber, S. (2017). Data, development, and growth. *Business and Politics*, 19(3), 397–423. <https://doi.org/10.1017/bap.2017.3>
- Webster, G., Creemers, R., Triolo, P., & Kania, E. (2017, August 1). *Full Translation: China's "New Generation Artificial Intelligence Development Plan" (2017)*. New America. <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/>
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. 2022 *ACM Conference on Fairness, Accountability, and Transparency*, 214–229. <https://doi.org/10.1145/3531146.3533088>
- Weiss, T. G. (2000). Governance, good governance and global governance: Conceptual and actual challenges. *Third World Quarterly*, 21(5), 795–814. <https://doi.org/10.1080/713701075>
- Woodward, R. (2009). *The Organisation for Economic Co-operation and Development (OECD)*. Routledge. <https://doi.org/10.4324/9780203875773>
- Zaidi, W., & Dafoe, A. (2021). *International Control of Powerful Technology: Lessons from the Baruch Plan for Nuclear Weapons* | GovAI. Future of Humanity Institute.

<https://www.governance.ai/research-paper/international-control-of-powerful-technology-lessons-from-the-baruch-plan-for-nuclear-weapons#>